2005 IEEE International
Symposium on Signal Processing
and Information Technology

# BIO101—EST Sequence Management and Annotation System

In-Yee Lee[1,2], Wen-Chang Lin[3], Jan-Ming Ho[2], Gu-Liang Wang[3], Ming-Syan Chen[1]

[1]*Department of Electrical Engineering, National Taiwan University, Taiwan*
[2] *Institute of Information Science, Academia Sinica, Taiwan*
[3]*Institute of Biomedical Sciences, Academia Sinica, Taiwan*
*E-mail: iylee@iis.sinica.edu.tw, wenlin@ibms.sinica.edu.tw,hoho@iis.sinica.edu.tw, guliang@ibms.sinica.edu.tw, mschen@cc.ee.ntu.edu.tw*

***Abstract*-***Laboratories worldwide are working on EST cDNA library and microarray projects, each project involving hundreds to hundreds of thousands of clones. To support the needs of individual laboratories, an efficient EST sequence management system is needed. Here we describe BIO101, a centralized approach to laboratory EST sequence management and functional annotation. The system includes a) laboratory-based authorization, project-focused and custom database management features; b) a pipeline with integrated tools for producing and clustering high-quality ESTs and for identifying homologues; c) a pipeline scheduling mechanism that emphasizes fairness; and d) collaborative manual and automatic functional annotation features. BIO101 is accessible to registered members at http://bio101.iis.sinica.edu.tw/.*

***Keywords****-expressed sequence tags, sequence management, annotation*

## 1. INTRODUCTION

This paper describes BIO101, a centralized approach to laboratory EST sequence management and functional annotation. Important advances in large-scale DNA sequencing have dramatically changed biological research. Expressed sequence tags (ESTs) are nucleotide sequences generated from "single-pass" cDNA libraries. dbEST—a database that contains sequence data and other information on ESTs from a number of organisms [4] is currently being used by a large number of biomedical researchers around the world for gene identification and characterization tasks, as well as for developing cDNA microarray platforms.

ESTs can provide significant additional functional, structural and evolutionary information. They are useful for identifying novel genes and for studying gene functions. Therefore, a lot of worldwide biological projects or laboratories continually produce ESTs for different researching tasks. Each project entails anywhere from a few hundred to hundreds of thousands of clones. The standard approach is to manage and analyze ESTs and cDNA clones manually—a labor-intensive task that can be performed more efficiently with the help of computer automation. Therefore, BIO101's first objective is to provide an efficient EST sequence management system, which is essential for individual projects to organize and process their EST sequences.

EST analysis entails preprocessing and sequence analysis. EST preprocessing involves the integration of various tools for generating high-quality sequences from machine-dependent trace files containing raw data. Trace files with different machine-dependent formats such as ABI, SCF, or MegaBASE are converted into nucleotide sequences [1], screened for low-complexity sequences and interspersed repeats[2], and cleaned of vectors and adaptors[3]. The primary purpose of processing the raw data is to generate high-quality EST sequences. When processing EST sequences, preprocessing and sequence analysis can be broken down into subtasks, all of which require significant human effort. BIO101's second objective is to address the automation of all tasks concerned with laboratory-based EST analysis.

The BIO101 architecture is shown in Figure 1. Automated pipelining (Fig. 1 center) allows for the scheduling of all tasks, including EST preprocessing, homologue identification, and information integration. To achieve its third objective—to provide a pipeline that emphasizes fairness—BIO101 incorporates a scheduling mechanism that allows for sequential processing of individual projects' tasks while achieving maximum global throughput. Project-based data management (Fig. 1 upper-right) makes it possible to work with data from different projects and to enforce role-based security policies.

BIO101's fourth objective is to provide collaborative manual and automated functional annotation features. Both automated and manual functional annotations entail more detailed EST sequence analyses that require additional data. Based on the assumption that similar sequences from different organisms share common biological characteristics and functions, a key strategy for annotating unknown EST sequences is to analyze information associated with similar sequences that have already been characterized. Information

[1] The Phred software reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base [18].
[2] RepeatMasker screens DNA sequence for interspersed repeats and low complexity DNA sequences [19].
[3] VecScreen is a system for quickly identifying segments of a nucleic acid sequence that maybe of vector origin.

sources include manual sequence descriptions and structured annotations that are usually found via hyperlinks to related information resources. Sources and their corresponding information extraction methods can be categorized as follows:

1. Free-formed descriptions [2] [20].

Lexical analyses and keyword-category dictionaries are used for the automated extraction of meaningful keywords. At least two problems exist with this method: keyword synonyms and definition ambiguity, and parallel proprietary dictionary development that often leads to duplicated efforts among various research groups.

2. Ontology frameworks [8] [11] [1] [27].

Due to vocabulary and format heterogeneity, functional information is difficult to manage electronically. It is therefore desirable to annotate uncharacterized sequences with the help of controlled vocabularies. Designers and researchers are increasingly turning to ontology frameworks for functional annotations, with the most popular being the Gene Ontology (GO) established by the Gene Ontology Consortium [24]. GO creates domain-specific vocabulary sets for describing molecular characteristics across various organisms. The three ontologies that constitute the GO term hierarchy are molecular functions, biological processes, and cellular components. Directed acyclic graphs (DAGs) are formed by GO terms and their "is-a" and "part-of" relations, meaning that GO provides both controlled and structured vocabularies.

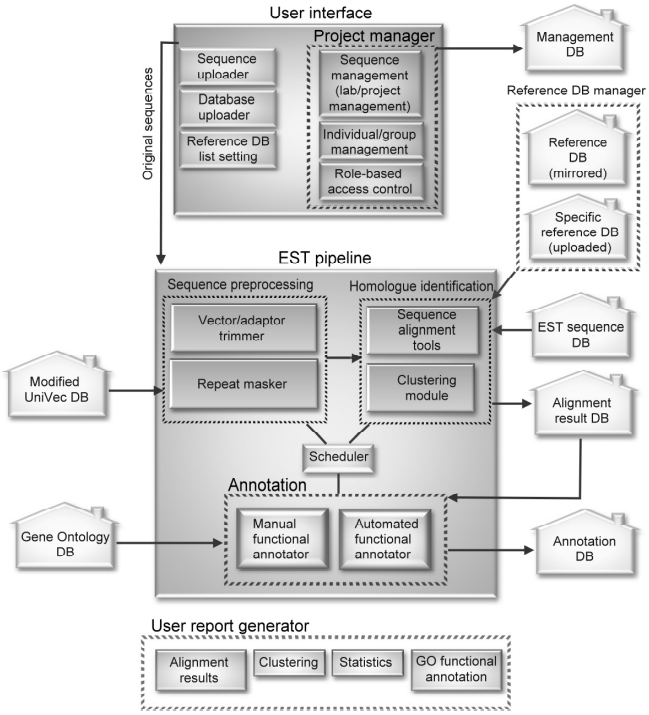


Fig. 1. System architecture and information flow.

BIO101 achieves its fourth objective by incorporating a GO-based annotation mechanism (Fig. 1 lower-center) for the automated (or manual, if desired) annotation of the biological properties of an unknown sequence.

The rest of this paper is organized as follows. Section 2 presents a brief survey of related works in the areas of EST pipeline design and annotation mechanisms. Section 3 describes the EST pipeline construction, Section 4 provides details on the annotation mechanism, and Section 5 discusses the sequence management features of BIO101. A conclusion is offered in Section 6.

## 2. RELATED WORK

BIO101 is associated with a combination of research on annotation mechanism and EST pipeline design. In this section, we will discuss other EST pipeline systems and compare various annotation mechanisms.

We noted that several EST pipeline systems share similar characteristics with our approach, with all of them providing information integration and tool automation for analyzing EST data. The systems we looked at include EST Pipeline System [28], ESTAP [15], ESTWeb [17], PipeOnline [2], ESTAnnotator [6], and ESTIMA [10]. Kumar's ESTIMA is unique in that it allows users to create multiple projects. The main difference in BIO101 is that BIO101 provides EST management features based on a project hierarchy that uses comprehensive authentication and authorization management and role-based security controls. Furthermore, our proposed approach makes use of a scheduling mechanism that enhances fairness and throughput maximization.

The two main annotation mechanism categories are manual and automated approaches. Frishman et al.'s PEDANT [5], designed for high-throughput analysis of large biological sequence sets, uses manual annotation as a means for refining its automated assignments. They addressed the problem of false annotations caused by a fully automatic approach, and suggested using manual annotation as a refinement mechanism.

Automatic annotation categories include statistical, machine learning, and information retrieval approaches. Leontovich et al. [14] proposed using an adaptive, statistical algorithm for automated annotation. Bazzan et al. [3] used machine learning techniques to generate keyword annotation rules, and evaluated their results by having experts manually validate their automatically generated rules.

Approaches based on information retrieval techniques include GeneQuiz [20], PipeOnline [2], and ESTAnnotator [6]. As a semi-automatic genome sequence analysis system, GeneQuiz performs functional assignment by generating a general functional classification dictionary through a process that involves a combination of human intervention

and automated keyword extraction. Keyword sequences are associated with functional class sets from the dictionary. Ayoubi et al.'s PipeOnline [2]'s functional annotation is based on their own NCBI protein-function dictionary, which is generated by matching EC numbers and keywords from the NCBI record description with functional definitions from the Metabolic Pathways Database (MPW) [21] dictionary.

GO-based information retrieval approaches include Xies et al.'s GO Engine [27] and Khan et al.'s GOFigure [9]. GO Engine serves as a computational platform for large-scale protein annotation that utilizes a mechanism based on a) sequence homologous having GO-annotated proteins and b) protein domain analysis. Text information analysis is used to increase accuracy. GOFigure [9] is a web-based tool for predicting GO terms; it constructs a minimum covering graph (MCG) that subsumes all the homologue annotations for an uncharacterized sequence. Each candidate term in the MCG is assigned a score for purposes of finding an optimal functional assignment.

For the purpose of providing uncharacterized sequence annotation, BIO101 provides both manual and automated approaches that make use of a novel mechanism for GO-based automated annotations (see Section 4.2).

## 3. EST PIPELINE

BIO101's pipeline engine (Fig. 1 center) comprises of three major components: an EST preprocessor, a homologue identifier, and an annotator. The scheduler schedules tasks to be processed by these tasks.

### *3.1 EST preprocessing*

BIO101 uses a set of public tools for vector stripping, contamination removal, and repeated masking (using RepeatMasker [19]). For vector stripping, we chose VecScreen [26], which is a system for quickly identifying segments of a nucleic acid sequence that maybe of vector origin. In conjunction with UniVec [25] database (which is a database containing vector sequences, adaptors, linkers and primers commonly used in the process of cloning cDNA or genomic DNA), a modified database (Fig. 1 left) that incorporates user-specified adaptors and vectors is developed for VecScreen [26]. All tools are interconnected within BIO101's pipeline, with program input (output) parsed and read from (written to) the backend MSSQL database. This makes it easy to integrate additional tools–for instance, Phred [18] for base calling and CAP3 for assembling tasks.

### *3.2 Homologue identification*

Annotating uncharacterized sequences using their homologue annotations involves the use of sequence alignment tools (Fig. 1 middle-right) and a set of reference databases (Fig. 1 upper-right). To support the proper alignment of various sequence types, we made use of the *BLAST* sequence alignment package [1]. The reference databases include the NCBI non-redundant amino acid/nucleic acid database, the NCBI RefSeq database [16], the TIGR Gene Indices which represents the most comprehensive, publicly available analysis of EST sequences [23], and the SWISS-PROT database, which is a curated protein sequence database that strives to provide a high level of annotation such as the description of protein function [22]. Since databases that cover various organisms and sequence types can be used simultaneously to produce integrated results, algorithms should be aimed at optimized selection of databases, alignment thresholds, and integrated functions [8]. BIO101 lets users choose the appropriate reference DBs and alignment thresholds. After an uncharacterized EST goes through a BIO101 pipeline, users can use the data generated by BIO101 for annotation purposes.

BIO101's reference DB manager (Fig. 1 upper-right) periodically mirrors the reference databases. In our opinion, it is impossible to mirror and centrally manage all available resources for all the organisms. Hence, we designed a web interface that allows users to upload specific databases to be integrated into the basic reference databases.

### *3.3 Scheduling*

In response to the time-consuming task of alignments between uncharacterized ESTs and whole sequences in reference databases, BIO101's scheduler (Fig. 1 middle) incorporates a pipeline scheduling mechanism that allows for concurrent use by multiple labs. The mechanism emphasizes fairness and throughput maximization (achieving the latter by executing consecutive programs without breaks). Fairness in terms of resource (CPU) usage is enforced by a system that maintains separate queues for short/light and long/heavy jobs. The system uses a first-come first-serve priority policy based on time of request, but prevents light-weight processes (jobs) from being locked out by heavy-weight ones. Jobs are designed as follows:

1. NEW_SUBMISSION: Each time when a user uploads sequences, BIO101 designates it as a NEW_SUBMISSION job and inserts it into the lightweight job queue.
2. NEW_REFDB: When a user either updates the list of reference databases or uploads new versions of private databases for a particular project, all associated sequences must be realigned using the newly-selected reference databases. Since this is a time-consuming task, BIO101

creates a NEW_REFDB job and inserts it into the heavyweight queue.

3. UPDATE_REFDB: Public databases are constantly being updated. Therefore we add a mechanism through which BIO101 downloads the newest version of the public databases using a pulling mechanism and replaces the old reference databases with the new ones. After such replacement, all sequences that have been aligned using the old database set must be re-aligned using the new set. In this case, BIO101 creates an UPDATE_REFDB job and inserts it into the heavyweight job queue. BIO101 performs this task both periodically and when no user-assigned tasks are pending.

The scheduler concurrently executes the jobs in the heavy-weight job queue and light-weight job queue, enforcing a first-come first-serve policy for each individual queue. This is achieved by using a multi-processor system, and allows for light-weight jobs to execute without having to wait for heavy-weight ones. This results in fair resource allocation.

## 4. ANNOTATION

BIO101's annotation engine (Fig. 1 bottom) supports two approaches for generating annotation reports—manual and automated. BIO101's manual approach supports a highly customizable annotation system that incorporates enhanced reporting and visualization features aimed at assisting the user with manual annotation, while its automated approach supports a system that is generally automated by using computational analysis.

### *4.1 Manual annotation*

In the manual approach, BIO101's annotation UI lets users select different BLAST programs and reference databases (Fig. 1 upper-left, *Reference DB list setting*). Using similarity scores ranked according to user-defined criteria (e.g., precision, alignment length and e-value), BIO101 generates reports containing the top three homologues.

As part of the manual approach, a graphical view of available information for an uncharacterized sequence (see Fig. 2) includes homologue descriptions, alignment results, and functional annotations, with the last item achieved by integrating the reference database with Gene Ontology (GO) functional annotation. Biologists can annotate sequences and set the confidence levels using the web-based UI. Higher confidence levels allow the pipeline to skip the alignment process for a particular EST. On the other hand, previous alignment can be tracked according to their histories. A fully computational approach is faster and provides clues for conducting further experiments to refine annotations. However, it is important to provide a manual annotation mechanism so as to reduce the likelihood for an incorrect annotation to be produced and to support the computational method [7].

To reduce the overhead caused by redundant EST sequences, our pipeline engine incorporates a clustering module that allows for self-clustering within individual projects (Fig. 1 middle). One source of overhead is the redundant annotations of multiple ESTs that are separate parts of the same gene. Clustering allows biologists to refine their further experiments and to select appropriate ESTs for microarray experiments. Clustering begins with a seed and performs one-against-all comparison to identify initial clusters. A sequence is assigned to a cluster if it overlaps with other members of that cluster. Overlapping criteria between two sequences are a) >50 residue overlap and b) >95% identity. When annotating a new EST, users refer to the annotations of its cluster members in order to select the most appropriate. Figure 2 demonstrates a manual annotation example.

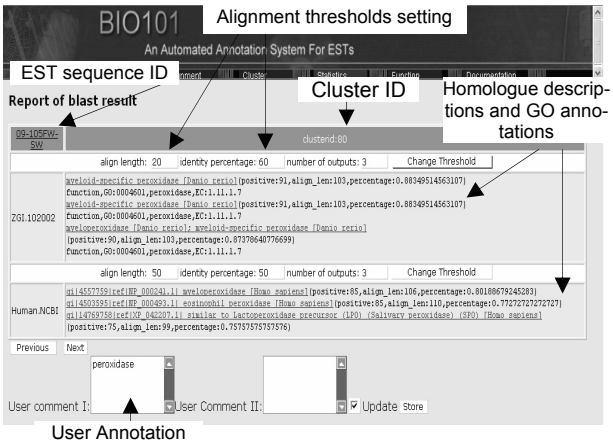


Fig. 2. A manual annotation example.

### *4.2 Automated annotation*

BIO101 also provides an automated mechanism that produces informative functional annotations. In a previous effort [11] [12] [13], we provided highly informative and comprehensive annotations using information revealed by the structured vocabularies of GO. After collecting the homologues' GO annotations, we use a novel ontology-based clustering algorithm for analyzing term distributions on a GO DAG [12] in order to identify groups of GO terms.

We illustrate this annotation process in Figure 3. A quantitative model assigns a score (confidence value) to each candidate term and selects the highest-scoring term as best describing a group [13]. Experiments showed that we can identify, with both high precision and recall, the most

informative terms that best describe the biological properties of a sequence [12] [13]. Figure 4 presents the annotation of a sample sequence; the GO term with the highest confidence value is selected as the best annotation for that sequence.

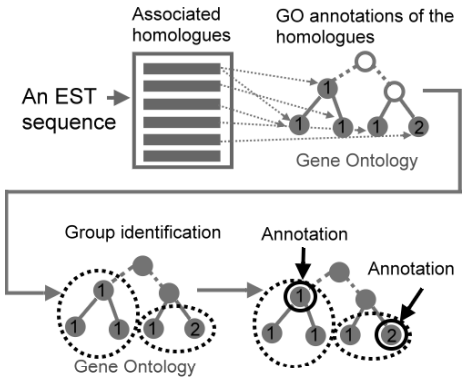


Fig. 3. Automated annotation process.

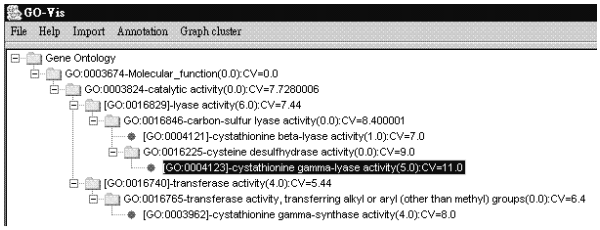


Fig. 4. Automated annotation of a sample sequence.

## 5. SEQUENCE MANAGEMENT SYSTEM DESIGN

We focused on centralized management processes in our BIO101 design in order to a) eliminate duplicated efforts among various labs in terms of software and database installation, tools searches, and other common processes; b) reduce human effort to sequentially process input; and c) reduce human effort in sequence data management.

BIO101's web-based user interface (Fig. 1 top, *User interface*) allows for uploading multiple sequences in a FASTA file and for pasting single sequences. The sequence manager provides sequence management within a hierarchical lab / project architecture (Fig. 5 top). Each project is associated with a unique list of basic and uploaded reference databases. Individual or group management and role-based access control can be created for individual projects (Fig. 5 bottom). Sequence access rights for authorized identities entail viewing, uploading, and annotating.

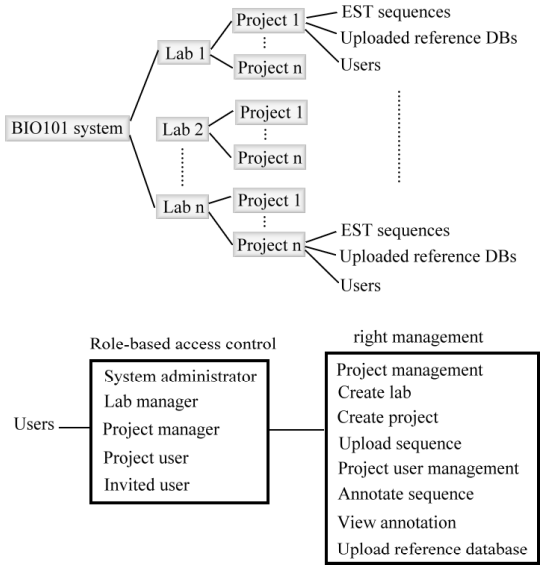


Fig. 5. The hierarchical management architecture.

## 6. CONCLUSION AND FUTURE WORK

Large-scale genome analysis and annotation has attracted a great deal of research attention. However, laboratory-scale systems are also important because lots of laboratories worldwide are involved in developing EST cDNA libraries. We designed BIO101 to support laboratory-scale requirements that include EST sequence management and processing and the generation of useful information in support of further biological experiments (e.g. microarray experiments). BIO101 provides automation of all processes involved in EST analysis and therefore reduces significant human effort. The automation is achieved by a) an automated pipeline and scheduling mechanism, b) laboratory-based management, and c) automated annotation. Furthermore, BIO101 also serves as an experimental platform for automated functional annotation research [12] [13]. We are interested in integrating various functional prediction methods into BIO101 to provide more clues for functional annotation of uncharacterized ESTs.

## ACKNOWLEDGEMENTS

The work was supported in part by the National Science Council of Taiwan, R.O.C., under Contracts NSC93-2752-E-002-006-PAE and NSC93-3112-B-001-018-Y.

## REFERENCE


[1] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ, "Basic local alignment search tool," *Journal of Molecular Biology*, 215, 403–410, 1990.

[2] Ayoubi P, Jin X, Leite1 S, Liu X, Martajaja1 J, Abduraham A, Wan Q, Yan W, Misawa1 E and Prade RA, "PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Research*, Vol. 30, No. 21, 2002.

[3] Bazzan AL, Engel PM, Schroeder LF and da Silva SC, "Automated annotation of keywords for proteins related to *mycoplasmataceae* using machine learning techniques**,"** *Bioinformatics* 2002 (18): S35-S43, 2002.

[4] Boguski MS, Lowe TM and Tolstoshev CM, "dbEST--database for expressed sequence tags," *Nature Genetics*, 1993 4(4):332-3, 1993.

[5] Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A and Mewes HW, "Functional and structural genomics using PEDANT," *Bioinformatics* 17, 44-57, 2001.

[6] Hotz-Wagenblatt A, Hankeln1 T, Ernst P, Karl-Heinz Glatting, Schmidt1 ER and Suhai S, "ESTAnnotator: a tool for high throughput EST annotation," *Nucleic Acids Research*, Vol. 31, No. 13, 2003.

[7] Kasukawa T, Furuno M, Nikaido I, Bono H, Hume DA, Bult C, Hill DP, Baldarelli R, Gough J, Kanapin A, Matsuda H, Schriml LM, Hayashizaki Y, Okazaki Y and Quackenbush J, "Development and evaluation of an automated annotation pipeline and cDNA annotation system," *Genome Research* 13: 1542-1551, 2003.

[8] Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I and Saito T, "Functional annotation of a full-length mouse cDNA collection," *Nature* 409, 685-690, 2001.

[9] Khan S, Situ G, Decker K and Schmidt CJ, "GoFigure: Automated Gene Ontology™ annotation," *Bioinformatics* 2003 19: 2484-2485, 2003.

[10] Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA, Liu L, "ESTIMA, a tool for EST management in a multi-project environment," *BMC Bioinformatics* 2004 Nov 4;5(1):176.

[11] Lee IY, Ho JM and Lin WC, "An algorithm for generating representative functional annotations based on Gene Ontology," *DEXA Workshops* 2003: 10-15.

[12] Lee IY, Ho JM, Chen MS, "CLUGO: A clustering algorithm for automated functional annotations based on Gene Ontology," *IEEE 5th International conference on Data Mining (ICDM)*, 2005.

[13] Lee IY, Ho JM, Chen MS, "GOMIT: A Generic and Adaptive Annotation Algorithm Based on Gene Ontology Term Distributions," *IEEE 5th Symposium on Bioinformatics and Bioengineering*, 40-48, 2005.

[14] Leontovich AM, Brodsky LI, Drachev VA and Nikolaev VK, "Adaptive algorithm of automated annotation," *Bioinformatics* (18): 838-844, 2002.

[15] Mao C, Cushman JC, May GD and Weller JW, "ESTAP—an automated system for the analysis of EST data," *Bioinformatics* 19: 1720-1722, 2003.

[16] NCBI RefSeq database: http://www.ncbi.nlm.nih.gov/RefSeq/index.html.

[17] Paquola AC, Nishyiama MY Jr, Reis EM, da Silva AM and Verjovski-Almeida S, "ESTWeb: bioinformatics services for EST sequencing projects," *Bioinformatics* 19(12):1587-1588, 2003.

[18] Phred: http://www.phrap.org/.

[19] RepeatMasker: http://repeatmasker.genome.washington.edu/.

[20] Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C and Sander C, "GeneQuiz: a workbench for sequence analysis," *ISMB* 2: 348-353, 1994.

[21] Selkov EJ, Grechkin Y, Mikhailova N and Salkov E, "MPW: the Metabolic Pathways Database," *Nucleic Acids Research*, Vol. 26, 43–45, 1998.

[22] SwissProt: http://tw.expasy.org/sprot/.

[23] TIGR Gene Indices: http://www.tigr.org/tdb/tgi/.

[24] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research* 11: 1425-1433, 2001.

[25] Univec: ftp://ftp.ncbi.nih.gov/pub/UniVec/.

[26] VecScreen: http://www.ncbi.nlm.nih.gov/VecScreen/.

[27] Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A and Mintz L, "Large-scale protein annotation through Gene Ontology," *Genome Research* 12, 785-794, 2002.

[28] Xu H, He L, Zhu Y, Huang W, Fang L, Tao L, Zhu Y, Cai L, Xu H, Zhang L, Xu H, Zhou Y, "EST pipeline system: detailed and automated EST data processing and mining," *Genomics Proteomics Bioinformatics* 1(3):236-42, 2003.